CS6923 Machine Learning, Fall 2018
Prof. Hellerstein, NYU School of Engineering

Homework 1 - Solutions (34 points)

Part I: Written Exercises (5+5+4+4+4 = 22 points)

1. (5 points)

   (a) (1 point)

   $$\begin{cases} 4 \cdot P(A) = P(B) \\ P(A) + P(B) = 1 \end{cases} \Longrightarrow \begin{cases} P(A) = \dfrac{1}{5} \\ P(B) = \dfrac{4}{5} \end{cases}$$

   (c) (2 points)

   Note: It's more convenient to do question c before question b here.

   Given Location A or B, we can calculate the probability of each rock belonging to that location using the standard Gaussian formula with given mean and std.

   $$\begin{aligned} p(x|A) &= (\frac{1}{\sqrt{2\pi}\sigma_1})^3 e^{-\frac{(x_1-\mu_1)^2+(x_2-\mu_1)^2+(x_3-\mu_1)^2}{2\sigma^2}} \\ &= (\frac{1}{\sqrt{2\pi}\cdot 1.6})^3 e^{-\frac{(9.3-9.2)^2+(8.8-9.2)^2+(9.8-9.2)^2}{2\cdot(1.6)^2}} \\ &\approx 0.013977 \end{aligned}$$

   $$\begin{aligned} p(x|B) &= (\frac{1}{\sqrt{2\pi}\sigma_2})^3 e^{-\frac{(x_2-\mu_2)^2+(x_2-\mu_2)^2+(x_3-\mu_2)^2}{2\sigma^2}} \\ &= (\frac{1}{\sqrt{2\pi}\cdot 1.2})^3 e^{-\frac{(9.3-9.6)^2+(8.8-9.6)^2+(9.8-9.6)^2}{2\cdot(1.2)^2}} \\ &\approx 0.0281237 \end{aligned}$$

   $p(x|A) < p(x|B)$, so the ML hypothesis is that the rocks are from Location B.

   (b) (2 points)

   $$\begin{aligned} P(A)p(x|A) &= \frac{1}{5} \cdot 0.013977 \\ &\approx 0.0027954 \end{aligned}$$

   $$\begin{aligned} P(B)p(x|B) &= \frac{4}{5} \cdot 0.0281237 \\ &\approx 0.02249896 \end{aligned}$$

   $$\begin{aligned} P(A|x) &= \frac{P(A)p(x|A)}{p(x)} \\ &= \frac{P(A)p(x|A)}{P(A)p(x|A) + P(B)p(x|B)} \\ &\approx 11.05\% \end{aligned}$$

   Posterior prob that the rock is from Location A = 11.05%

   Note: we compute $P(x) = P(A)p(x|A) + P(B)p(x|B)$.

2. (5 points)

   From the question we can obtain that:

$$
\begin{array}{cc}
P(D) = 0.0002 & P\left(\overline{D}\right) = 0.9998 \\
P\left(pos_1|\overline{D}\right) = 0.15 & P\left(notpos_1|\overline{D}\right) = 0.85 \\
P(pos_1|D) = 0.9 & P(notpos_1|D) = 0.1 \\
P\left(pos_2|\overline{D}\right) = 0.05 & P\left(notpos_2|\overline{D}\right) = 0.95 \\
P(pos_2|D) = 0.97 & P(notpos_2|D) = 0.03
\end{array}
$$

   (b) (1 point)

   $P(pos_1|D) = 0.9$

   $P(pos_1|\overline{D}) = 0.15$

   Since $P(pos_1|D) > P(pos_1|\overline{D})$, so the ML hypothesis is that the patient has the disease for ML hypothesis.

   (a) (2 points)

$$
\begin{aligned}
P(D|pos_1) &= P(D)P(pos_1|D)/P(pos_1) \\
&= 0.0002 \cdot 0.9/P(pos_1) \\
&= 0.00018/P(pos_1)
\end{aligned}
$$

$$
\begin{aligned}
P(\overline{D}|pos_1) &= P(\overline{D})P(pos_1|\overline{D})/P(pos_1) \\
&= 0.9998 \cdot 0.15/P(pos_1) \\
&= 0.14997/P(pos_1)
\end{aligned}
$$

   Since $P(\overline{D}|pos_1) > P(D|pos_1)$, so the MAP hypothesis is that the patient does not have the disease.

   (c) (2 points)

   $x = \{pos_1, pos_2\}$

$$
\begin{aligned}
P(x|D) \cdot P(D) &= P(pos_1|D) \cdot P(pos_2|D) \cdot P(D) \\
&= 0.9 \cdot 0.97 \cdot 0.0002 \\
&= 0.0001746
\end{aligned}
$$

$$
\begin{aligned}
P\left(x|\overline{D}\right) \cdot P\left(\overline{D}\right) &= P\left(pos_1|\overline{D}\right) \cdot P\left(pos_2|\overline{D}\right) \cdot P\left(\overline{D}\right) \\
&= 0.15 \cdot 0.05 \cdot 0.9998 \\
&= 0.0074985
\end{aligned}
$$

$$
\begin{aligned}
P(D|x) &= \frac{P(x|D) \cdot P(D)}{P(x)} \\
&= \frac{P(x|D) \cdot P(D)}{P(x|D) \cdot P(D) + P\left(x|\overline{D}\right) \cdot P\left(\overline{D}\right)} \\
&= 0.0227548
\end{aligned}
$$

3. (4 points)

   (a) (1 point)

   $P(TTHTTH|\theta) = \theta^2(1-\theta)^4$

(b) (1 point)
$$\log P(TTHTTH|\theta) = \log(\theta^2(1-\theta)^4)$$
$$= 2\log\theta + 4\log(1-\theta)$$

(c) (2 points)
Let $f(\theta) = \log P(TTHTTH|\theta) = 2\log\theta + 4\log(1-\theta)$.
We need the value of θ for which $f(\theta)$ is maximum i.e $f'(\theta) = 0$ and $f''(\theta) < 0$.

$$\frac{\partial}{\partial\theta} logP(TTHTTH|\theta) = -\frac{4}{1-\theta} + \frac{2}{\theta} = 0$$

Solving, $\theta = \frac{1}{3}$.
Therefore, $argmax_\theta P(TTHTTH|\theta) = argmax_\theta \log P(TTHTTH|\theta) = \dfrac{1}{3}$
Ideally, we should verify that this value indeed maximizes $f(\theta)$ by checking if $f''(\theta) < 0$

4. (4 points)

(a) (1 point)
ML estimate for θ_1 is $\dfrac{2}{5}$.

(b) (1 point)
Using add-1 smoothing. It is $\dfrac{2+1}{5+3\times 1} = \dfrac{3}{8}$.

(c) (2 points)
$P(\theta_1|X) = \frac{P(\theta_1)\cdot P(X|\theta_1)}{P(X)}$
To get MAP estimate of $\theta_1 \equiv argmax_{\theta_1} P(\theta_1|X)$
$\Rightarrow argmax_{\theta_1} P(X|\theta_1)$
$= argmax_{\theta_1} 6\theta_1(1-\theta_1)\theta_1^6(1-\theta_1)^{N-t}$
(Since there are t is in X) Get log.
Thus to get $argmax_{\theta_1} log6 + (t+1)log_{\theta_1} + (N-t+1)\cdot log(1-\theta_1)$
Get Deviations to tell at which point it get max value
$$\Rightarrow argmax_{\theta_1} \frac{t+1}{\theta_1} - \frac{N-t+1}{1-\theta_1}$$
$$\Rightarrow argmax_{\theta_1} \frac{t+1-(N+2)\cdot\theta_1}{\theta_1(1-\theta_1)}$$
Since $\theta \in [0,1]$, so it's easy to tell that from $(0, \frac{t+1}{N+2})$ it increses. Then it decreases between $(\frac{t+1}{N+2}, 1)$.
Therefore, when $\theta_1 = \frac{t+1}{N+2}$, we get MAP estimate of θ_1.

5. (4 points)

(a) (1 point)
$$P(x_1 = Low|+) = \frac{1+0.3}{2+0.3\times 3} = \frac{1.3}{2.9}$$

$$P(x_2 = Yes|+) = \frac{0+0.3}{2+0.3\times 2} = \frac{0.3}{2.6}$$
$$P(x_3 = Green|+) = \frac{1+0.3}{2+0.3\times 2} = \frac{1.3}{2.6} = \frac{1}{2}$$
$$P(x_1 = Low|-) = \frac{1+0.3}{3+0.3\times 3} = \frac{1}{3}$$
$$P(x_2 = Yes|-) = \frac{2+0.3}{3+0.3\times 2} = \frac{2.3}{3.6}$$
$$P(x_3 = Green|-) = \frac{2+0.3}{3+0.3\times 2} = \frac{2.3}{3.6}$$

(b) (1 point)

$$P(x|+) = \frac{1.3}{2.9} \times \frac{0.3}{2.6} \times \frac{1}{2} = \frac{0.3}{11.6} = 0.026$$
$$P(x|-) = \frac{1}{3} \times \frac{2.3}{3.6} \times \frac{2.3}{3.6} = \frac{5.29}{38.8} = 0.136$$

(c) (1 point)

Since $P(x|-) > P(x|+)$, for the example $x = [Low, Yes, Green]$, ML label is "$-$".

(d) (1 point)

$$P(+|x) = \frac{P(+)\cdot P(x|+)}{P(x)} = \frac{1}{P(x)} \cdot \left(\frac{2}{5} \cdot \frac{0.3}{11.6}\right) = \frac{1}{P(x)} \cdot 0.0104$$
$$P(-|x) = \frac{P(-)\cdot P(x|-)}{P(x)} = \frac{1}{P(x)} \cdot \left(\frac{3}{5} \cdot \frac{5.29}{38.8}\right) = \frac{1}{P(x)} \cdot 0.0816$$

So MAP label is "$-$".

Part II: Programming and Questions (12 points)

1. (1 point) 0.4018006002
2. (1 point) 0.5981993998
3. (2 points) 97.2091286307, 36369.9911126
4. (2 points) 0.0484258639911, 0.0883056032571
5. (1 point) 0, 0, 0, 0, 0
6. (1 point) 0, 0, 0, 0, 0
7. (1 point) 20%
8. (1 point) 59%
9. (2 points)

   You can get full credit by reasonable answers. The idea is to get you to think. Here are examples of such answers:

   For the conditional independence assumption you could say: No, I don't think the conditional independence assumption is reasonable, because I think that e.g. in spam emails, certain attributes would be correlated. Consider the attributes:

4. char_freq_!

percentage of total characters in the email that are equal to !

8. capital_run_length_longest

length of longest uninterrupted sequence of capital letters

9. capital_run_length_total

total number of capital letters in the email

I think that in spam emails, having a lot of exclamation points in the email would be more likely if the email has lots of capital letters. I say this because of my experience with Spam, and the fact that people who use lots of capital letters tend to be more emotional/shouting so they are more likely to use exclamation marks. So I think Attribute 4 would be positively correlated with Attribute 8 and/or Attribute 9, in the spam emails.

One could also run tests of statistical independence on attributes in e.g. the spam emails. See e.g. `http://www.stat.wmich.edu/s216/book/node112.html`. I don't expect you to know this – but some students may have learned to do something like this in a statistics course.

One thing to note: The conditional independence assumption does NOT say that all attributes are independent. It says that attributes are independent GIVEN THE CLASS. So it's important when considering the conditional independence assumption to focus just on the spam emails, or just on the non-spam emails, but not both together.

Regarding the Gaussian assumption:

Probably the easiest way to show that the Gaussian assumption is inappropriate is to consider a particular class C and attribute x_i, and to make a 1-d plot of the value of x_i in all examples in class C (you could then turn this into a histogram, if you want). Visually, you can then see whether the plot seems to follow a Gaussian, or if it follows some other shape. For example, if the plot looks something like

x x x x x x x x x x x x x

you might say that it looks more like $p(x_i|C)$ is uniform within some interval, rather than Gaussian.

(There are also statistical tests to determine whether data is likely to have come from a Gaussian distribution, but I doubt anyone will use these.)

I suppose you could try to make a common sense argument in English about why some particular attribute wouldn't follow a Gaussian distribution – e.g. you could say that in non-spam emails, most emails will not have more than one capital letter in a row, because they will only use capital letters at the start of a name or sentence. So in non-spam emails, the most frequent value of attribute 8 will be one, there will be fewer non-spam emails where the value will be 0 (no capital letters), and there will be very few emails where the value of attribute 8 will be greater than one (and if so, the value might be two or three or four). This wouldn't be Gaussian because the most common value is 1, but the frequency of the other values would not be symmetric around that value.

For full credit, you should talk about BOTH assumptions.

You can also get full credit for saying the the assumptions ARE reasonable, provided that they give a reasonable explanation. That is, they should either give a graph or some statistics to show that the assumption seems to be true (e.g. the plot looks Gaussian) or they should give some sort of intuitive argument.